**AirBNB price prediction Rio De Janeiro**

**Ana Pjević i Milica Medić**

1. Motivation

Motivation for this project is automatically offering reasonable price for apartments, houses, etc in the city of Rio de Janeiro.

2. Research questions

Problem that is solved in this project is finding appropriate price for AirBNB accommodation

automatically. This is very helpful due to not many renters being experienced in finding appropriate price for their accommodation. Dataset is .csv file downloaded from <http://insideairbnb.com/get-the-data.html> site in Rio de Janeiro section. Our dataset is the first file in the list of files related to Rio.

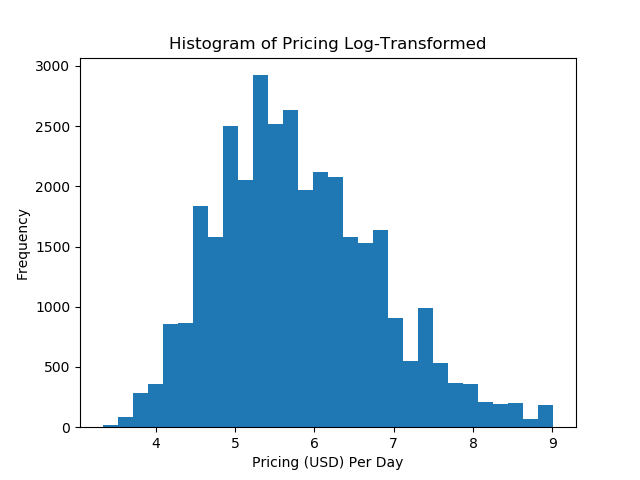
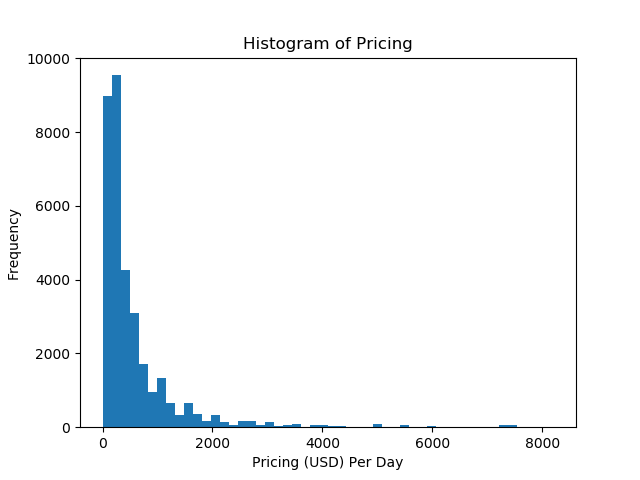
3. Related work

One of the papers we analyzed is *Airbnb Price Estimation*, Liubov Nikolenko, Hoormazd Rezaei, Pouya Rezazadeh. They used data preprocessing techniques and feature selection before training models. Methods used in this research are: Ridge Regression, K-means Clustering with Ridge Regression, Support Vector Regression, Neural Network and Gradient Boost Tree Ensemble.

4. Methodology

First, we preprocessed our data. Due to irrelevance, redundancy and inconvenience for processing of many columns (59 specifically), for example id, description, *host\_total\_listings\_count*, etc we decided to drop that columns. Also, we dropped columns with too much NA values in their fields. Later, we decided to fill NA values. For columns that have too little NA values, we dropped rows containing that values. For others, we used median values. Other part of preprocessing was transforming data in order to have values that we can work with. For example, we counted members of list values in *host\_verifications* column fields and added column *host\_verifications\_count* with that values. Finally, we encoded categorical data using *One hot encoder* and made 238 columns in total.

Data distributions are given in the following pictures:

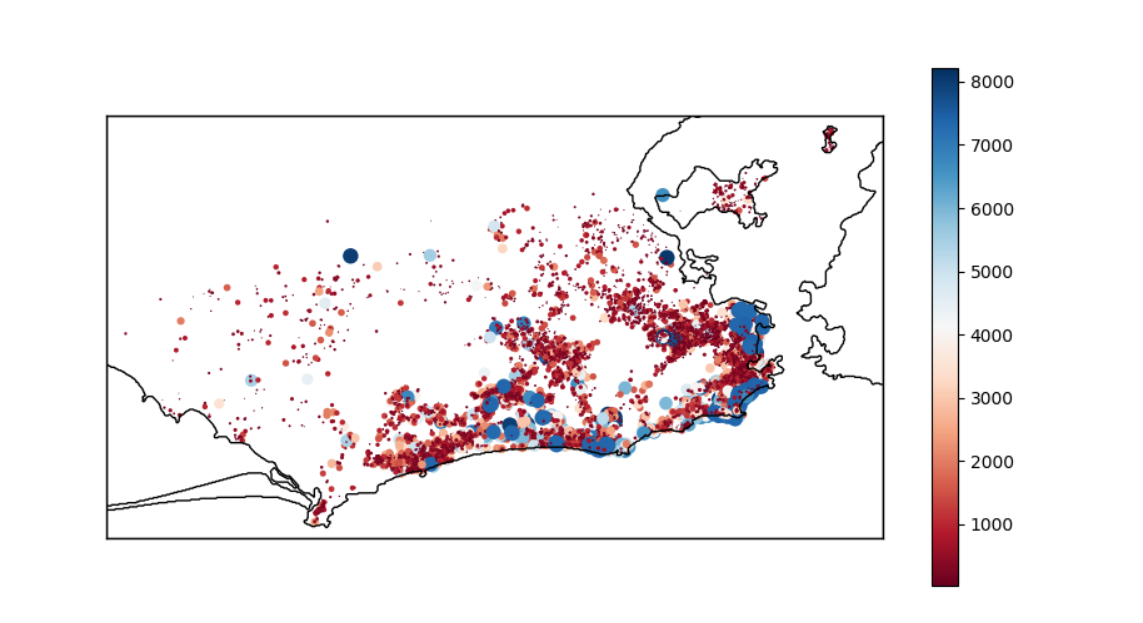
: 

Pictures 1 and 2: Data distributions

In order to process data more successfully we normalized data from every column that is not categorical. Specifically, we user Z Score normalization. Outliers are also removed.

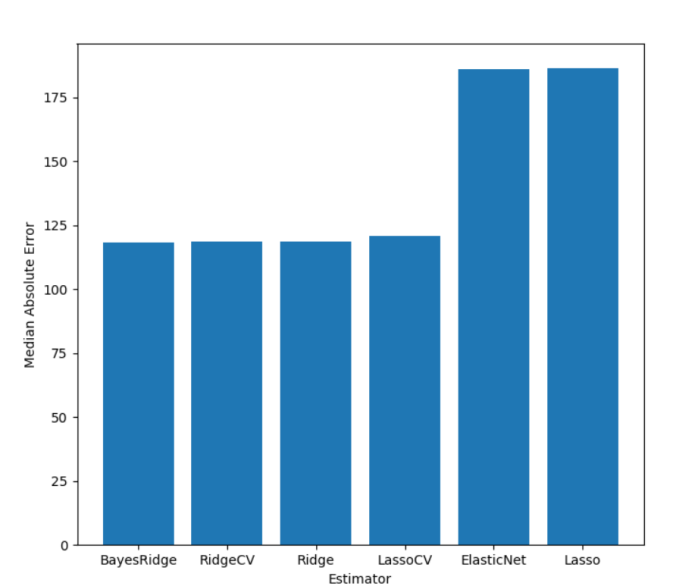
For solving the problem we used many models and compared the results. We used: Ridge, Lasso, Ridge and Lasso with cross validation (CV), Elastic Net, Bayesian Ridge models. Best result was: Bayesian Ridge model with MAE (Median Absolute Error) = 118.20318885705888.

5. Discussion



Picture 3: Density of data depending on longitude and latitude

First and very important part of the procedure was to see the data to intuitively have review of the problem. We started from that, and processed data to get to the core of the problem: predicting price. In some models we used cross-validation, dividing data into 3 parts: training, validation and test. In others we used train and test data in ratio 65:35. Then we trained our models described in subtitle above, got to error analysis and got the results on the picture:



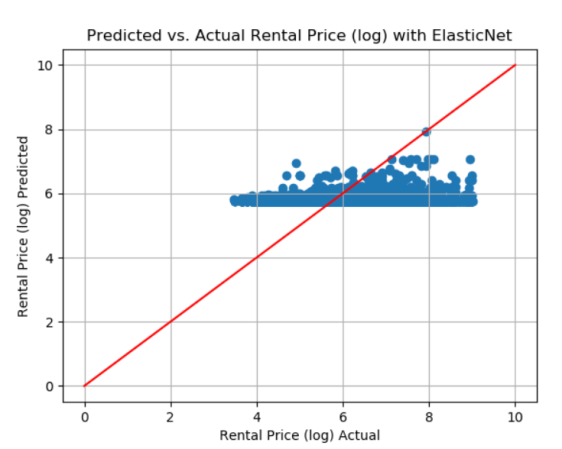
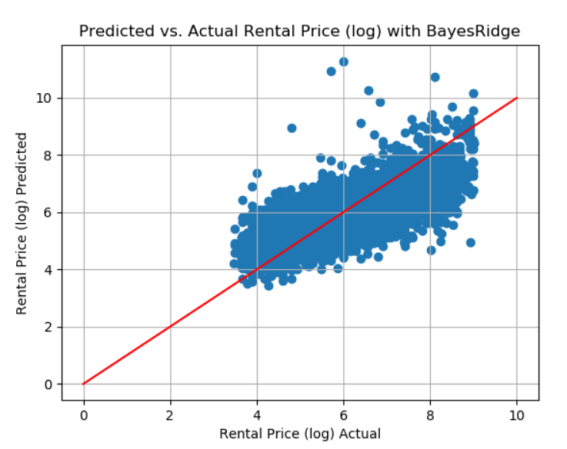
Picture 4: Error analysis data

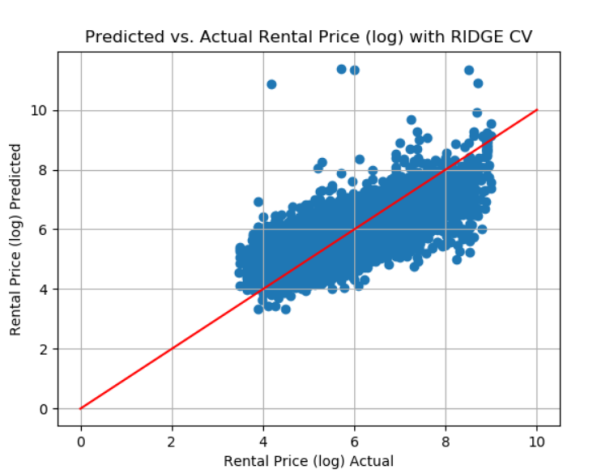
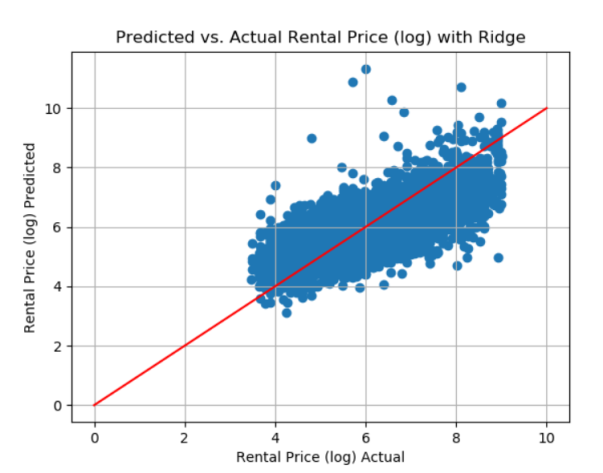
As seen, Bayesian Ridge showed the best results, showing just a little better result than Ridge CV, Ridge and Lasso CV.

Exact results are :

* Model Ridge Results: 118.55232615479082
* Model RidgeCV Results: 118.54830903959936
* Model Lasso Results: 186.5774825799918
* Model LassoCV Results: 120.78008968237825
* Model ElasticNet Results: 186.00613378685125
* Model BayesRidge Results: 118.20318885705888

Some examples of predicted vs. actual rental prices (the more data are following the trend of the red line, the more precise the model is):





6. References

References :

* [1] <http://cs229.stanford.edu/proj2016/poster/WuYu-HousePricePrediction-poster.pdf?fbclid=IwAR09CWWzmy38t5OV0v4GHdB30U3D5GFvEruZvaVtZ6dnlJaVF_9zMrf4bx4>
* [2] <https://www.dataquest.io/blog/machine-learning-tutorial/?fbclid=IwAR2egtjGlh0U5zuysSII-wpdPhphCra3xfDcJfLK3tYK7H1siAfmfyHRl7M>
* [3]<http://cs229.stanford.edu/proj2018/report/96.pdf?fbclid=IwAR2DKDkJXmFQrkTT3-ok8ddVVUgdKBP24gM9AGEFPrV2bOfvQrbEd78THwg>
* [4] <https://towardsdatascience.com/digging-into-airbnb-data-reviews-sentiments-superhosts-and-prices-prediction-part1-6c80ccb26c6a>